TILBURG UNIVERSITY

Tilburg University

Optimization of Periodic Polling Systems with Non-Preemptive, Time-Limited Service

Blanc, J.P.C.

Publication date:
1996

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
Blanc, J. P. C. (1996). *Optimization of Periodic Polling Systems with Non-Preemptive, Time-Limited Service*. (CentER Discussion Paper; Vol. 1996-63). Operations research.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 05. May. 2023

Optimization of periodic polling systems with non-preemptive, time-limited service

J. P. C. Blanc
Tilburg University, CentER for Economic Research,
P.O. Box 90153, 5000 LE Tilburg, The Netherlands.

July 5, 1996

Abstract

This paper is devoted to polling systems with periodic visit orders and non-preemptive, time-limited service. Performance measures of these systems are evaluated with the aid of the power-series algorithm (PSA), which is a flexible technique for computing performance measures for multi-queue systems with a Markovian structure of the queue-length process. For application of the PSA it is necessary to approximate the constant time limits by Erlang distributed random variables. The PSA can be extended to compute derivatives of performance measures with respect to parameters of a system. This allows for optimization of cost functions with respect to the mean values of the time limits by means of gradient methods. Several properties of the optimal time limits are revealed by the numerical solution of various optimization problems, some of them including constraints on the time limits.

Keywords: Power-series algorithm; cost of waiting; polling table.

1 Introduction

In many communication, production and other systems several types of jobs compete for access to a single service facility, e.g., a communication channel or a machine. Such systems are often modeled as polling systems. These are multi-queue systems with a single server, who attends to the jobs in the various queues according to some visit rule and some service rules. The visit rule determines the order in which the queues are visited by the server. The service rules determine the number of services that the server is allowed to perform during the subsequent visits to the queues. The choice of these rules may partly be limited by physical constraints, but can otherwise be used to control the quality of service provided to each of the job types.
Polling systems are generally hard to analyse. For some systems it is possible to derive sets of linear equations that determine the moments of the waiting time distributions. Examples are systems with periodic visit orders and with exhaustive or gated service, cf., e.g., Baker & Rubin [1]. Some two-queue models can be solved analytically for a larger class of service disciplines, but to obtain numerical data from these solutions may require substantial effort, cf. Coffman et al. [7] which deals with a two-queue system with exponentially distributed timers. In most other cases, performance measures can only be approximated by numerical techniques based on the solution of balance equations for state probabilities, or estimated by simulation. The power-series algorithm (PSA) is one of the available methods. It requires a Markov representation of the queueing process,

possibly with the aid of some supplementary variables. It is based on power-series expansions of the state probabilities in terms of a parameter of a system for (recursively) solving the global balance equations satisfied by these probabilities. It is a flexible method which is applicable to a wide class of multi-queue/multi-server models, with Markovian Arrival Processes (MAPs) and phase-type (PH) service time distributions. The PSA is also suitable for optimization purposes, since it allows the computation of derivatives of performance measures with respect to system parameters and control variables. For moderately sized systems, the PSA favourably compares with simulation and numerical methods based on truncation of the state space. This is mainly so because the PSA involves recursive schemes and allows the application of the so-called ϵ-algorithm which strongly improves the convergence of the power series, cf. [2]. Since the memory requirements rapidly grow with the number of queues, the PSA can only produce accurate results for systems with a limited number of queues. The main contribution of the PSA lies in studying the interaction between queues on a reduced scale and in developing and testing approximations of performance measures and optimal values of control variables for systems of a larger size. The paper [3] reviews the PSA in its generality. The applicability and complexity of the PSA for polling systems with various visit and service rules have been discussed in [2]. The computation of derivatives with the aid of the PSA has been described in a general context in Blanc & Van der Mei [5]. The latter paper considers as an example the problem of optimizing cost functions for polling systems with respect to the parameters of the service rules, in that case so-called Bernoulli schedules. In Borst et al. [6], the PSA has been used to test the quality of several approximative approaches for determining the optimal job limits in cyclic polling systems.

An alternative technique for computing performance measures of polling systems is the discrete Fourrier transform method. It is based on relations for the generating function of the queue length distribution at various imbedded time instants. In [9], Leung applies the discrete Fourrier transform method to cyclic polling systems with time-limited service. This method allows for general service and switching time distributions. However, the time limits have to be approximated by exponentially distributed timers. In the present paper, we develop the PSA for this type of polling models. The service and switching time distributions have to be approximated by phase type distributions. In fact, we will use Cox distributions because they are somewhat easier to describe and implement. But the method can be extended to general phase type distributions, cf., e.g., Van den Hout & Blanc [10]. On the other hand, the results of Leung [9] will be extended in several directions. Firstly, we will incorporate Erlang distributed timers in order to approximate the constant time limits more closely. Secondly, our model will allow for general periodic visit orders. Finally, we will consider the problem of choosing the values of the time limits such as to minimize some cost function, with or without restrictions on the time limits. For this purpose, the computation scheme of the PSA includes recursions for the coefficients of derivatives of performance measures with respect to the mean time limits.

The paper is organized as follows. In section 2 the polling model will be described in more detail, and the necessary notations will be introduced. Section 3 contains the global balance equations for the queue-length process extended with several supplementary variables. The recurrence relations of the computation scheme of the PSA for the state probabilities are given in section 4. Here, the influence of the number of phases of the Erlang distributions for the timers is illustrated. Section 5 extends the recursive scheme of the PSA to the computation of derivatives of the state probabilities with respect to the transitions rates of the timers. Section 6 deals with optimization of the cost of waiting with respect to the mean values of the time limits, possibly with constraints on individual time limits and on the sum of all time limits. Some of the results are compared with the optimal

values of job limits for similar systems. Finally, the conclusions are summarized in section 7.

2 Description of the model

The polling system consists of s queues and a single server. Jobs arrive at queue j according to a Poisson process with rate λ_j, $j = 1, \ldots, s$. The superposition of the arrival processes at the various queues is a Poisson process with rate $\Lambda \doteq \sum_{j=1}^{s} \lambda_j$. Each queue may contain an unbounded number of jobs. At each queue jobs are served in order of arrival. Service times of jobs arriving at queue j are assumed to be Coxian distributed with means β_j, $j = 1, \ldots, s$. The Laplace-Stieltjes transform (LST) of the service time distribution for jobs at queue j is:

$$\beta_j(\zeta) \doteq \sum_{\phi=1}^{\Psi_j} \pi_{j,\phi} \prod_{\psi=1}^{\phi} \frac{\mu_{j,\psi}}{\mu_{j,\psi} + \zeta}, \quad \Re\zeta \geq 0; \tag{2.1}$$

i.e., the distribution consists of Ψ_j exponential phases, with probability $\pi_{j,\phi}$ a service consists of consecutive phases $\phi, \phi - 1, \ldots, 1$, $\phi = 1, \ldots, \Psi_j$, and the transition rate at phase ψ is $\mu_{j,\psi}$, $\psi = 1, \ldots, \Psi_j$, $j = 1, \ldots, s$. The load ρ_j offered at queue j, $j = 1, \ldots, s$, and the total offered load ρ to the system are defined by

$$\rho_j \doteq \lambda_j \beta_j, \qquad \rho \doteq \sum_{j=1}^{s} \rho_j. \tag{2.2}$$

The server visits the queues in a fixed periodic order which is determined by a polling table of length P. This table will be described by a mapping $\ell : \{1, \ldots, P\} \to \{1, \ldots, s\}$. It will be assumed throughout that each station occurs at least once on the table, and that P has been chosen as small as possible, given a fixed visit order. Further, the convention $\ell(0) = \ell(P)$ will be needed and used. The number of services which may be performed during the hth visit of the server during a tour according to the polling table $\{\ell(1), \ldots, \ell(P)\}$ is determined by a time limit τ_h, $h = 1, \ldots, P$. As long as this time limit has not expired the server is allowed to start new services during a visit. A visit to a queue ends either when a service is completed and the current visit already lasts longer than the time limit or when the queue is or becomes empty. In practice, the time limits will be constant. However, in order to construct a Markov process these time limits will be approximated by random variables with Erlang distributions. The number of phases of the Erlang distribution of the time limit for the hth visit of the server during a tour according to the polling table will be denoted by Γ_h, and the transition rates at the phases are $\gamma_h \doteq \Gamma_h / \tau_h$, so that the mean value of the Erlang distributed time limit is τ_h, $h = 1, \ldots, P$.
The times the server needs for switching from queue $\ell(h-1)$ to queue $\ell(h)$ are assumed to be Coxian distributed with means δ_h, $h = 1, \ldots, P$. The LST of the switching time distribution between queue $\ell(h-1)$ and queue $\ell(h)$ is: for $h = 1, \ldots, P$,

$$\delta_h(\zeta) \doteq \sum_{\phi=1}^{\Omega_h} \omega_{h,\phi} \prod_{\psi=1}^{\phi} \frac{\nu_{h,\psi}}{\nu_{h,\psi} + \zeta}, \quad \Re\zeta \geq 0; \tag{2.3}$$

here, the parameters have similar interpretations as those of the service time distributions. For example, if the switching time distribution between queue $\ell(h-1)$ and queue $\ell(h)$ is Erlang with Ω_h phases for some $h \in \{1, \ldots, P\}$, then $\omega_{h,\Omega_h} = 1$, $\omega_{h,\phi} = 0$ for $\phi = 1, \ldots, \Omega_h - 1$, and $\nu_{h,\psi} = \nu_{h,1}$, $\psi = 2, \ldots, \Omega_h$. The total mean switching time of the server during a tour according to the polling table will be denoted by $\Delta \doteq \sum_{h=1}^{P} \delta_h$.

From the general result on stability of periodic polling systems in Fricker & Jaïbi [8] it follows that the present system is stable iff

$$\rho + \Delta \max_{j=1,\ldots,s} \{\lambda_j / G_j\} < 1; \tag{2.4}$$

here, G_j denotes the mean of the maximal number of jobs that can be served at station j during a tour of the server according to the polling table, $j = 1, \ldots, s$. If some station j is visited more than once during a tour then G_j is the sum of the mean of the maximal numbers of jobs that can be served during the various visits to this station, $j = 1, \ldots, s$. The expression for the mean of the maximal numbers of jobs that can be served during a visit to a station with a time limit is not as simple as that for stations with a job limit or a Bernoulli schedule. It will be discussed in more detail in appendix A.

3 The balance equations

It will be assumed throughout this paper that the polling systems are in steady state. The random variable N_j will indicate the number of jobs present at queue j, $j = 1, \ldots, s$. Beside the vector of random variables $\mathbf{N} \doteq (N_1, \ldots, N_s)$ several supplementary variables are needed to obtain a Markov process. The supplementary variable H will indicate the queue to which the server is switching or to which the server is attending. The supplementary variable Z will indicate the status of the server. More precisely, $Z = 0$ will indicate that the server is switching and $Z = \kappa$ will indicate that the server is serving jobs while the timer is in phase κ, $\kappa = 1, \ldots, \Gamma_H + 1$; here, $Z = \Gamma_H + 1$ indicates that the timer has already expired during the current visit. The supplementary variable Φ will indicate the actual phase of the current switching time or service time. The Markov process $(\mathbf{N}, H, Z, \Phi)$ is assumed to be stable. In order to formulate the balance equations for this Markov process we will use the indicator function $I_{\{C\}}$ taking the values 0 (if C is false) or 1 (if C is true), and the unit vectors $\mathbf{e}_j$, $j = 1, \ldots, s$, in $\mathbb{N}^s$. The balance equations for the probabilities of states in which the server is switching are, for $\mathbf{n} \in \mathbb{N}^s$, $h = 1, \ldots, P$, $\phi = 1, \ldots, \Omega_h$,

$$\begin{aligned}
[\Lambda + \nu_{h,\phi}] p(\mathbf{n}, h, 0, \phi) &= \sum_{j=1}^{s} \lambda_j I_{\{n_j \geq 1\}} p(\mathbf{n} - \mathbf{e}_j, h, 0, \phi) + \nu_{h,\phi+1} I_{\{\phi < \Omega_h\}} p(\mathbf{n}, h, 0, \phi + 1) \\
&+ \nu_{h-1,1} \omega_{h,\phi} I_{\{n_{\ell(h-1)} = 0\}} p(\mathbf{n}, h-1, 0, 1) + \mu_{\ell(h-1),1} \omega_{h,\phi} p(\mathbf{n} + \mathbf{e}_{\ell(h-1)}, h-1, \Gamma_{h-1} + 1, 1) \\
&+ \mu_{\ell(h-1),1} \omega_{h,\phi} I_{\{n_{\ell(h-1)} = 0\}} \sum_{\kappa=1}^{\Gamma_{h-1}} p(\mathbf{n} + \mathbf{e}_{\ell(h-1)}, h-1, \kappa, 1).
\end{aligned} \tag{3.1}$$

The first term at the righthand side stands for transitions caused by an arrival of a job at one of the queues. The second term stands for a phase transition in the switching time. The third term describes a transition from a switch to queue $\ell(h-1)$ to a switch to queue $\ell(h)$; such a transition can only occur if queue $\ell(h-1)$ is empty. The fourth and fifth term describe a transition from a last service at queue $\ell(h-1)$ to a switch to queue $\ell(h)$; in the fourth term, the end of the visit is due to the expiration of the timer at queue $\ell(h-1)$, in the fifth one to the exemption of queue $\ell(h-1)$.
The balance equations for the probabilities of states in which the server is serving jobs are, for $\mathbf{n} \in \mathbb{N}^s$, $h = 1, \ldots, P$, $n_{\ell(h)} \geq 1$, $\kappa = 1, \ldots, \Gamma_h + 1$, $\phi = 1, \ldots, \Psi_{\ell(h)}$,

$$[\Lambda + \mu_{\ell(h),\phi} + \gamma_h I_{\{\kappa \leq \Gamma_h\}}] p(\mathbf{n}, h, \kappa, \phi) = \sum_{j=1}^{s} \lambda_j I_{\{n_j \geq 1\}} p(\mathbf{n} - \mathbf{e}_j, h, \kappa, \phi)$$

$$+ \mu_{\ell(h),\phi+1} I_{\{\phi<\Psi_{\ell(h)}\}} p(\mathbf{n},h,\kappa,\phi+1) + \nu_{h,1}\pi_{\ell(h),\phi} I_{\{\kappa=1\}} p(\mathbf{n},h,0,1)$$
$$+ \gamma_h I_{\{\kappa\geq 2\}} p(\mathbf{n},h,\kappa-1,\phi) + \mu_{\ell(h),1}\pi_{\ell(h),\phi} I_{\{\kappa\leq\Gamma_h\}} p(\mathbf{n}+\mathbf{e}_h,h,\kappa,1). \tag{3.2}$$

The first term at the righthand side stands for transitions caused by an arrival of a job at one of the queues. The second term stands for a phase transition in the service time. The third term describes a transition from a switch to queue $\ell(h)$ to the first service at queue $\ell(h)$ (the timer starts in phase $\kappa = 1$). The fourth term describes a phase transition of the timer. The fifth term describes a transition from one service at queue $\ell(h)$ to another service at queue $\ell(h)$; such a transition can only occur if the timer had not expired before the new service started, i.e., if $\kappa \leq \Gamma_h$. Note that for all $\mathbf{n} \in \mathbb{N}^s$, $h = 1,\ldots,P$, $\kappa = 1,\ldots,\Gamma_h+1$, $\phi = 1,\ldots,\Psi_{\ell(h)}$,

$$p(\mathbf{n},h,\kappa,\phi) = 0, \quad \text{if } n_{\ell(h)} = 0, \tag{3.3}$$

because the server cannot be serving a job at a queue which is empty.
Finally, it holds by the law of total probability that

$$\sum_{n_1=0}^{\infty} \cdots \sum_{n_s=0}^{\infty} \sum_{h=1}^{P} \left[\sum_{\phi=1}^{\Omega_h} p(\mathbf{n},h,0,\phi) + \sum_{\kappa=1}^{\Gamma_h+1} \sum_{\phi=1}^{\Psi_{\ell(h)}} p(\mathbf{n},h,\kappa,\phi) \right] = 1. \tag{3.4}$$

4 The power-series algorithm

We introduce power-series expansions of the state probabilities as functions of the total offered load ρ: for all $\mathbf{n} \in \mathbb{N}^s$, $h = 1,\ldots,P$, $\kappa = 0,1,\ldots,\Gamma_h+1$, $\phi = 1,\ldots,\Omega_h$ if $\kappa = 0$, $\phi = 1,\ldots,\Psi_{\ell(h)}$ if $\kappa = 1,\ldots,\Gamma_h+1$,

$$p(\mathbf{n},h,\kappa,\phi) = \rho^{|\mathbf{n}|} \sum_{k=0}^{\infty} \rho^k b(k;\mathbf{n},h,\kappa,\phi). \tag{4.1}$$

Here and below, we use the notation $|\mathbf{n}| \doteq n_1 + \cdots + n_s$. We write $\lambda_j = a_j\rho$, $j = 1,\ldots,s$, and $\Lambda = A\rho$ to obtain a parametrization of the model as a function of ρ, cf. Blanc [3]. These expressions and the expansions (4.1) are substituted into the equations (3.1) and (3.2) for the state probabilities. Equating coefficients of corresponding powers of ρ on both sides of these equations leads to relations for the coefficients of the power-series expansions of the state probabilities. The recurrence relations for the coefficients of the probabilities of states in which the server is switching are, for $k = 0,1,2,\ldots$, $\mathbf{n} \in \mathbb{N}^s$, $h = 1,\ldots,P$, $\phi = 1,\ldots,\Omega_h$,

$$\begin{aligned}
\nu_{h,\phi} b(k;\mathbf{n},h,0,\phi) &= \sum_{j=1}^{s} a_j I_{\{n_j\geq 1\}} b(k;\mathbf{n}-\mathbf{e}_j,h,0,\phi) - A I_{\{k\geq 1\}} b(k-1;\mathbf{n},h,0,\phi) \\
&+ \nu_{h,\phi+1} I_{\{\phi<\Omega_h\}} b(k;\mathbf{n},h,0,\phi+1) + \nu_{h-1,1}\omega_{h,\phi} I_{\{n_{\ell(h-1)}=0\}} b(k;\mathbf{n},h-1,0,1) \\
&+ \mu_{\ell(h-1),1}\omega_{h,\phi} I_{\{k\geq 1\}} b(k-1;\mathbf{n}+\mathbf{e}_{\ell(h-1)},h-1,\Gamma_{h-1}+1,1) \\
&+ \mu_{\ell(h-1),1}\omega_{h,\phi} I_{\{n_{\ell(h-1)}=0\}} I_{\{k\geq 1\}} \sum_{\kappa=1}^{\Gamma_{h-1}} b(k-1;\mathbf{n}+\mathbf{e}_{\ell(h-1)},h-1,\kappa,1).
\end{aligned} \tag{4.2}$$

The recurrence relations for the coefficients of the probabilities of states in which the server is serving jobs are, for $k = 0,1,2,\ldots$, $\mathbf{n} \in \mathbb{N}^s$, $h = 1,\ldots,P$, $n_{\ell(h)} \geq 1$, $\kappa = 1,\ldots,\Gamma_h+1$, $\phi = 1,\ldots,\Psi_{\ell(h)}$,

$$[\mu_{\ell(h),\phi} + \gamma_h I_{\{\kappa\leq\Gamma_h\}}] b(k;\mathbf{n},h,\kappa,\phi)$$

$$= \sum_{j=1}^{s} a_j I_{\{n_j \geq 1\}} b(k; \mathbf{n} - \mathbf{e}_j, h, \kappa, \phi) - A I_{\{k \geq 1\}} b(k-1; \mathbf{n}, h, \kappa, \phi)$$
$$+ \mu_{\ell(h),\phi+1} I_{\{\phi < \Psi_{\ell(h)}\}} b(k; \mathbf{n}, h, \kappa, \phi+1) + \nu_{h,1} \pi_{\ell(h),\phi} I_{\{\kappa=1\}} b(k; \mathbf{n}, h, 0, 1)$$
$$+ \gamma_h I_{\{\kappa \geq 2\}} b(k; \mathbf{n}, h, \kappa-1, \phi) + \mu_{\ell(h),1} \pi_{\ell(h),\phi} I_{\{\kappa \leq \Gamma_h, k \geq 1\}} b(k-1; \mathbf{n} + \mathbf{e}_h, h, \kappa, 1). \quad (4.3)$$

The relations (4.2) and (4.3) can be used to compute the coefficients of the power-series expansions of the state probabilities in a mainly recursive manner, cf. [2], [3], using the same order of computation as for polling systems with job-limited service. The only term which may prevent recursive computation is the term with $b(k; \mathbf{n}, h-1, 0, 1)$ in (4.2). This term only occurs if $n_{\ell(h-1)} = 0$. The only case in which the coefficients can not be computed recursively is the case $\mathbf{n} = \mathbf{0}$; this is the only situation in which the server can make a complete tour along the queues without any change in the values of $\mathbf{N}$. In the case $\mathbf{n} = \mathbf{0}$ equation (4.2) reduces to: for $k = 0, 1, 2, \ldots$, $h = 1, \ldots, P$, $\phi = 1, \ldots, \Omega_h$,

$$\nu_{h,\phi} b(k; \mathbf{0}, h, 0, \phi) = \nu_{h,\phi+1} I_{\{\phi < \Omega_h\}} b(k; \mathbf{0}, h, 0, \phi+1) + \nu_{h-1,1} \omega_{h,\phi} b(k; \mathbf{0}, h-1, 0, 1)$$
$$+ I_{\{k \geq 1\}} \left[\mu_{\ell(h-1),1} \omega_{h,\phi} \sum_{\kappa=1}^{\Gamma_{h-1}+1} b(k-1; \mathbf{n} + \mathbf{e}_{\ell(h-1)}, h-1, \kappa, 1) - A b(k-1; \mathbf{0}, h, 0, \phi) \right]. \quad (4.4)$$

This forms, for each fixed k, $k = 0, 1, 2, \ldots$, a dependent set of equations for the coefficients $b(k; \mathbf{0}, h, 0, \phi)$, $h = 1, \ldots, P$, $\phi = 1, \ldots, \Omega_h$. These sets of equations have the same structure as those which have been encountered in periodic polling models with Bernoulli schedules, cf. [2], although the general structure of the relations (4.2) and (4.3) is quite different from that of the latter model. These sets of equations can be solved together with the following relations which follow from the law of total probability (3.4) and relation (3.3): for $k = 0$,

$$\sum_{h=1}^{P} \sum_{\phi=1}^{\Omega_h} b(0; \mathbf{0}, h, 0, \phi) = 1; \quad (4.5)$$

respectively for $k = 1, 2, \ldots$,

$$\sum_{h=1}^{P} \sum_{\phi=1}^{\Omega_h} b(k; \mathbf{0}, h, 0, \phi)$$
$$= - \sum_{1 \leq |\mathbf{n}| \leq k} \cdots \sum \sum_{h=1}^{P} \left[\sum_{\phi=1}^{\Omega_h} b(k - |\mathbf{n}|; \mathbf{n}, h, 0, \phi) + \sum_{\kappa=1}^{\Gamma_h+1} \sum_{\phi=1}^{\Psi_{\ell(h)}} b(k - |\mathbf{n}|; \mathbf{n}, h, \kappa, \phi) \right]. \quad (4.6)$$

The coefficients of the power-series expansions of the moments are obtained from those of the state probabilities in a straightforward manner, cf. [3]. The moments of the waiting time distributions for jobs at the various queues, assuming service in order of arrival, follow from the moments of the marginal queue-length distribution through a general relationship between the generating function of the queue-length distribution and the Laplace-Stieltjes transform of the waiting time in M/G/1-type systems, cf. e.g. [2]. The waiting time of a job at queue j is denoted by W_j, $j = 1, \ldots, s$, and W stands for the waiting time of an arbitrary job. The standard deviation of a random variable X will be denoted by $\sigma\{X\}$.

In table 1 the influence of the number of phases of the Erlang distributions of the timers is illustrated for an example taken from Leung [9]. The model consists of three queues and service is in cyclic

Γ_1	Γ_2	Γ_3	$\mathcal{V}$	$E\{W_1\}$	$E\{W_2\}$	$E\{W_3\}$	$E\{W\}$	$\sigma\{W_1\}$	$\sigma\{W_2\}$	$\sigma\{W_3\}$	$\sigma\{W\}$
1	1	1	18	2.200	5.724	5.802	3.625	2.612	8.664	8.749	6.121
2	1	1	19	2.056	5.903	5.984	3.611	2.385	8.873	8.959	6.232
4	1	1	21	1.978	6.003	6.088	3.605	2.251	8.994	9.084	6.300
8	1	1	25	1.939	6.054	6.142	3.603	2.177	9.060	9.153	6.338
16	1	1	33	1.921	6.079	6.169	3.602	2.138	9.095	9.190	6.358
32	1	1	49	1.912	6.091	6.182	3.602	2.119	9.114	9.210	6.368
64	1	1	81	1.907	6.097	6.188	3.601	2.109	9.123	9.219	6.373
2	2	2	21	2.060	5.855	5.955	3.598	2.370	8.804	8.905	6.187
4	2	2	23	1.979	5.960	6.065	3.592	2.228	8.930	9.036	6.258
8	2	2	27	1.938	6.013	6.122	3.590	2.150	8.999	9.109	6.298
16	2	2	35	1.919	6.039	6.150	3.589	2.110	9.035	9.148	6.318
32	2	2	51	1.910	6.052	6.164	3.589	2.089	9.054	9.168	6.329
64	2	2	83	1.905	6.058	6.170	3.589	2.079	9.064	9.178	6.334
4	4	4	27	1.978	5.937	6.056	3.585	2.213	8.893	9.010	6.234
8	4	4	31	1.936	5.992	6.116	3.583	2.132	8.964	9.085	6.275
16	4	4	39	1.916	6.019	6.144	3.582	2.090	9.001	9.125	6.296
32	4	4	55	1.906	6.032	6.158	3.582	2.069	9.020	9.146	6.307
64	4	4	87	1.902	6.038	6.165	3.582	2.058	9.030	9.156	6.313
8	8	8	39	1.934	5.981	6.113	3.579	2.121	8.945	9.073	6.263
16	8	8	47	1.914	6.008	6.142	3.578	2.079	8.983	9.113	6.285
32	8	8	63	1.904	6.022	6.156	3.578	2.057	9.002	9.134	6.296
64	8	8	95	1.899	6.028	6.163	3.578	2.047	9.012	9.145	6.301
128	8	8	159	1.897	6.031	6.167	3.578	2.041	9.017	9.150	6.304
128	16	16	175	1.895	6.026	6.167	3.576	2.035	9.007	9.144	6.299
128	32	32	207	1.894	6.024	6.167	3.575	2.031	9.003	9.141	6.296
160	40	40	255	1.894	6.023	6.168	3.574	2.029	9.003	9.143	6.296
120	60	60	255	1.894	6.023	6.167	3.574	2.029	9.003	9.143	6.296

Table 1: Three-queue model with varying number of phases of the timers.

$\Gamma_1^{(a)}$	$\Gamma_2^{(a)}$	$\Gamma_3^{(a)}$	$\Gamma_1^{(b)}$	$\Gamma_2^{(b)}$	$\Gamma_3^{(b)}$	$E\{W_1^{(\mathrm{Det})}\}$	$E\{W_2^{(\mathrm{Det})}\}$	$E\{W_3^{(\mathrm{Det})}\}$	$E\{W^{(\mathrm{Det})}\}$
1	1	1	2	2	2	1.920	5.986	6.108	3.571
2	2	2	4	4	4	1.896	6.019	6.157	3.572
2	1	1	4	2	2	1.902	6.017	6.146	3.573
4	2	2	8	4	4	1.893	6.024	6.167	3.574
4	1	1	8	2	2	1.898	6.023	6.156	3.575

Table 2: Estimation of the performance of the system with constant timers.

order. The arrival rates are $\lambda_1 = 0.6$, $\lambda_2 = \lambda_3 = 0.2$, the service times are exponential with means $\beta_j = 0.8$, $j = 1,2,3$, and the switching times are Erlang E_4 distributed with means $\delta_j = 0.05$, $j = 1,2,3$. The quantity $\mathcal{V}$ denotes the size of the supplementary space, cf. the last factor in (5.7). The mean values of the timers are $\tau_1 = 23.2$, $\tau_2 = \tau_3 = 1.6$, so that on the average 30 jobs can be served during a visit of the server to queue 1 and 3 jobs during visits to both queue 2 and 3. We have selected this model, because Leung [9] found the largest differences in mean waiting times between systems with exponential timers and constant timers in this example. In [9] this model has been solved with constant switching times and exponential timers ($\Gamma_j = 1$, $j = 1,2,3$); the reported mean waiting times are $E\{W_1\} = 2.198$, $E\{W_2\} = 5.713$ and $E\{W_3\} = 5.792$. Note that our results with E_4 distributed switching times and exponential timers are close to these values. Leung [9] also reports simulation results for this model with constant switching times and constant timers: $E\{W_1\} = 1.884 \pm 0.020$, $E\{W_2\} = 5.989 \pm 0.132$ and $E\{W_3\} = 6.163 \pm 0.148$. Our results with E_4 distributed switching times and Erlang distributed timers show that a rather large number of phases of the corresponding Erlang distribution is required to obtain a close approximation for the performance measures of a similar system but with constant timers, when the time limit at a queue is relatively large compared to the mean service time. However, it is shown in table 2 that simple linear extrapolations in the squared coefficients of variation of the Erlang distributions of the timers yield good approximations for the performance measures of the system with constant timers, based on the evaluation of two systems with Erlang distributed timers with only a few phases. Note that the squared coefficient of variation of an Erlang distribution with Γ phases is $1/\Gamma$, while that of a constant is 0. In the table, each row shows two sets of numbers of phases of the Erlang distributions of the timers, indicated by (a) and (b). The performance measures of the system with constant timers are estimated from those with the Erlang timers by the simple extrapolation

$$E\{X^{(\mathrm{Det})}\} \approx E\{X^{(b)}\} - [E\{X^{(a)}\} - E\{X^{(b)}\}] = 2E\{X^{(b)}\} - E\{X^{(a)}\}. \tag{4.7}$$

Here, X is some performance measure, $X^{(\mathrm{Det})}$ indicates the version with deterministic timers, and $X^{(a)}$ and $X^{(b)}$ stand for the versions with Erlang timers.

5 Derivatives with the PSA

For optimization of a performance measure with respect to real-valued parameters of a system it is useful to be able to compute derivatives of the performance measure as function of these parameters. Then, optimization techniques as the conjugate gradient method can be used to determine optimal values of these parameters with respect to some objective function. Computation of derivatives may also be useful to study the sensitivity of performance measures for changes in system parameters. For the present model, consider derivatives of the state probabilities with respect to the transition rates of the timers γ_h, $h = 1,\ldots,P$. It can be shown that these derivatives possess power-series expansions of the form, cf. (4.1): for all $\mathbf{n} \in \mathbb{N}^s$, $h,r = 1,\ldots,P$, $\kappa = 0,1,\ldots,\Gamma_h+1$, $\phi = 1,\ldots,\Omega_h$ if $\kappa = 0$, $\phi = 1,\ldots,\Psi_{\ell(h)}$ if $\kappa = 1,\ldots,\Gamma_h+1$,

$$\frac{\partial}{\partial\gamma_r} p(\mathbf{n},h,\kappa,\phi) = \rho^{|\mathbf{n}|} \sum_{k=0}^{\infty} \rho^k b_r(k;\mathbf{n},h,\kappa,\phi). \tag{5.1}$$

Taking derivatives in relations (4.2) and (4.3) with respect to γ_h, $h = 1,\ldots,P$, leads to a further set of recursions. The recurrence relations for the coefficients of the derivatives of the probabilities of

states in which the server is switching are, for $k = 0, 1, 2, \ldots$, $\mathbf{n} \in \mathbb{N}^s$, $h, r = 1, \ldots, P$, $\phi = 1, \ldots, \Omega_h$,

$$\begin{aligned}
\nu_{h,\phi} b_r(k;\mathbf{n},h,0,\phi) &= \sum_{j=1}^{s} a_j I_{\{n_j \geq 1\}} b_r(k;\mathbf{n}-\mathbf{e}_j,h,0,\phi) - A I_{\{k\geq 1\}} b_r(k-1;\mathbf{n},h,0,\phi) \\
&+ \nu_{h,\phi+1} I_{\{\phi<\Omega_h\}} b_r(k;\mathbf{n},h,0,\phi+1) + \nu_{h-1,1}\omega_{h,\phi} I_{\{n_{\ell(h-1)}=0\}} b_r(k;\mathbf{n},h-1,0,1) \\
&+ \mu_{\ell(h-1),1}\omega_{h,\phi} I_{\{k\geq 1\}} b_r(k-1;\mathbf{n}+\mathbf{e}_{\ell(h-1)},h-1,\Gamma_{h-1}+1,1) \\
&+ \mu_{\ell(h-1),1}\omega_{h,\phi} I_{\{n_{\ell(h-1)}=0\}} I_{\{k\geq 1\}} \sum_{\kappa=1}^{\Gamma_{h-1}} b_r(k-1;\mathbf{n}+\mathbf{e}_{\ell(h-1)},h-1,\kappa,1). \qquad (5.2)
\end{aligned}$$

The recurrence relations for the coefficients of the derivatives of the probabilities of states in which the server is serving jobs are, for $k = 0, 1, 2, \ldots$, $\mathbf{n} \in \mathbb{N}^s$, $h, r = 1, \ldots, P$, $n_{\ell(h)} \geq 1$, $\kappa = 1, \ldots, \Gamma_h+1$, $\phi = 1, \ldots, \Psi_{\ell(h)}$,

$$\begin{aligned}
&[\mu_{\ell(h),\phi} + \gamma_h I_{\{\kappa\leq\Gamma_h\}}] b_r(k;\mathbf{n},h,\kappa,\phi) \\
&\quad = \sum_{j=1}^{s} a_j I_{\{n_j\geq 1\}} b_r(k;\mathbf{n}-\mathbf{e}_j,h,\kappa,\phi) - A I_{\{k\geq 1\}} b_r(k-1;\mathbf{n},h,\kappa,\phi) \\
&\quad + \mu_{\ell(h),\phi+1} I_{\{\phi<\Psi_{\ell(h)}\}} b_r(k;\mathbf{n},h,\kappa,\phi+1) + \nu_{h,1}\pi_{\ell(h),\phi} I_{\{\kappa=1\}} b_r(k;\mathbf{n},h,0,1) \\
&\quad + \gamma_h I_{\{\kappa\geq 2\}} b_r(k;\mathbf{n},h,\kappa-1,\phi) + \mu_{\ell(h),1}\pi_{\ell(h),\phi} I_{\{\kappa\leq\Gamma_h\}} I_{\{k\geq 1\}} b_r(k-1;\mathbf{n}+\mathbf{e}_h,h,\kappa,1) \\
&\quad + I_{\{r=h\}}[I_{\{\kappa\geq 2\}} b(k;\mathbf{n},h,\kappa-1,\phi) - I_{\{\kappa\leq\Gamma_h\}} b(k;\mathbf{n},h,\kappa,\phi)]. \qquad (5.3)
\end{aligned}$$

Notice that (5.2) has exactly the same structure as (4.2), and, hence, reduces for $\mathbf{n} = \mathbf{0}$ to a set of equations similar to (4.4). The law of total probability leads in a similar way to: for $k = 0$, $r = 1, \ldots, P$,

$$\sum_{h=1}^{P} \sum_{\phi=1}^{\Omega_h} b_r(0;\mathbf{0},h,0,\phi) = 0; \qquad (5.4)$$

respectively for $k = 1, 2, \ldots$, $r = 1, \ldots, P$,

$$\begin{aligned}
&\sum_{h=1}^{P} \sum_{\phi=1}^{\Omega_h} b_r(k;\mathbf{0},h,0,\phi) \\
&\quad = -\sum_{1\leq|\mathbf{n}|\leq k} \cdots \sum \sum_{h=1}^{P} \left[\sum_{\phi=1}^{\Omega_h} b_r(k-|\mathbf{n}|;\mathbf{n},h,0,\phi) + \sum_{\kappa=1}^{\Gamma_h+1} \sum_{\phi=1}^{\Psi_{\ell(h)}} b_r(k-|\mathbf{n}|;\mathbf{n},h,\kappa,\phi) \right]. \qquad (5.5)
\end{aligned}$$

From (5.4) and (5.2) it follows that $b_r(0;\mathbf{0},h,0,\phi) = 0$ for $r = 1, \ldots, P$, $h = 1, \ldots, P$, $\phi = 1, \ldots, \Omega_h$. Note that the coefficients $b_r(0;\mathbf{0},h,\kappa,\phi)$ do not vanish for $\kappa \geq 1$, in contrast with those of similar polling models but with Bernoulli schedules as service discipline, cf. [3].
By means of (5.2), (5.3) and (5.5) the coefficients $b_r(k;\mathbf{n},h,\kappa,\phi)$ can be computed recursively, but only in conjunction with the coefficients $b(k;\mathbf{n},h,\kappa,\phi)$. Derivatives of performance measures with respect to the (mean) time limits can be computed from the above: for all $\mathbf{n} \in \mathbb{N}^s$, $h, r = 1, \ldots, P$, $\kappa = 0, 1, \ldots, \Gamma_h + 1$, $\phi = 1, \ldots, \Omega_h$ if $\kappa = 0$, $\phi = 1, \ldots, \Psi_{\ell(h)}$ if $\kappa = 1, \ldots, \Gamma_h + 1$,

$$\frac{\partial}{\partial\tau_r} p(\mathbf{n},h,\kappa,\phi) = -\frac{\Gamma_r}{\tau_r^2} \frac{\partial}{\partial\gamma_r} p(\mathbf{n},h,\kappa,\phi). \qquad (5.6)$$

The number of coefficients which have to be computed in order to determine the power-series expansions of performance measures and their derivatives with respect to R parameters (in the present model, mean time limits for the duration of visits to various queues) up to the Mth power of ρ is

$$(R+1)\binom{M+s+1}{s+1}\sum_{h=1}^{P}[\Omega_h+(\Gamma_h+1)\Psi_{\ell(h)}]. \tag{5.7}$$

The above computation scheme is readily extended to the computation of second order derivatives but the latter requires still more additional storage space.

6 Optimization of the time limits

Consider the following optimization problem with the time limits as decision variables:

$$\min_{\tau_1,\ldots,\tau_P} C \doteq \sum_{j=1}^{s} c_j E\{W_j\}, \tag{6.1}$$

subject to

$$L_h \le \tau_h \le U_h, \quad h=1,\ldots,P; \tag{6.2}$$

$$\sum_{h=1}^{P}\tau_h \le B. \tag{6.3}$$

The coefficients c_j in the objective function indicate the relative cost of waiting for jobs at station j, $j=1,\ldots,s$. By taking $c_j \doteq \lambda_j/\Lambda$, $j=1,\ldots,s$, the objective becomes minimization of the overall mean waiting time.

Note that $\tau_h = 0$ implies $\gamma_h = \infty$, $h=1,\ldots,P$. If a time limit vanishes no job would ever be served during the corresponding visits to a queue. We have taken $L_h = 10^{-6}$, $h=1,\ldots,P$, in all examples to prevent that γ_h becomes too large. A very small but positive value of a time limit means in fact that the server is allowed to serve exactly one job during each visit to the corresponding queue.

On the other hand, a large value of a time limit means that the corresponding queue is served exhaustively, i.e., until it becomes empty. In the cases that we indicate that the optimal time limit is infinite the optimization procedure stopped at some finite value of that time limit because the derivative of the cost function with respect to that time limit became too small. We have also evaluated the cost function in those cases with a much larger value of the time limit, and have found no significant differences in costs.

Tables 3 and 4 show the unconstrained optimal time limits as function of the load for a cyclic three-queue system with the following parameters: arrival rates $\lambda_1 = \frac{2}{3}\rho$, $\lambda_2 = \lambda_3 = \frac{1}{3}\rho$, exponentially (table 3) respectively Erlang E_2 (table 4) distributed service times with means $\beta_1 = 1.0$, $\beta_2 = \beta_3 = 0.5$, Erlang E_2 distributed switching times with means $\delta_j = 0.1$, $j=1,2,3$, and cost factors $c_1 = 0.8$, $c_2 = c_3 = 0.1$. The optimal time limits are shown for the case of exponentially distributed timers as well as for the case of Erlang E_2 respectively Erlang E_4 distributed timers at all queues. In the cases of exponentially and Erlang E_2 distributed timers (with minimal cost indicated by C_{M} respectively C_{E_2}) the system has also been evaluated with the same values of the mean time limits, but with Erlang E_4 distributed timers (cost C_{E_4}). The results in the tables illustrate that the cost functions are rather flat near their minima, and that optimization is less sensitive to the number of phases of the timers than evaluation of performance measures. Note the resemblance of

	Exponential timers					Erlang-2 timers					Erlang-4 timers			
ρ	τ_1^*	τ_2^*	τ_3^*	C_{M}	C_{E_4}	τ_1^*	τ_2^*	τ_3^*	C_{E_2}	C_{E_4}	τ_1^*	τ_2^*	τ_3^*	C_{E_4}
0.3	∞	∞	∞	0.548	0.548	∞	∞	∞	0.548	0.548	∞	∞	∞	0.548
0.4	∞	2.53	3.16	0.748	0.747	∞	1.72	1.87	0.747	0.746	∞	1.43	1.57	0.746
0.5	∞	1.60	1.70	1.020	1.017	∞	1.28	1.43	1.018	1.017	∞	1.18	1.30	1.017
0.6	∞	1.45	1.48	1.416	1.410	∞	1.26	1.31	1.412	1.410	∞	1.18	1.24	1.410
0.7	∞	1.66	1.65	2.060	2.046	∞	1.47	1.49	2.050	2.046	∞	1.37	1.42	2.046
0.8	∞	2.41	2.25	3.314	3.281	∞	2.09	2.06	3.290	3.279	∞	1.95	1.97	3.279
0.9	∞	4.93	4.35	6.987	6.894	∞	4.22	4.05	6.923	6.888	∞	3.88	3.94	6.887

Table 3: Optimal time limits as function of the load: exponential service.

	Exponential timers					Erlang-2 timers					Erlang-4 timers			
ρ	τ_1^*	τ_2^*	τ_3^*	C_{M}	C_{E_4}	τ_1^*	τ_2^*	τ_3^*	C_{E_2}	C_{E_4}	τ_1^*	τ_2^*	τ_3^*	C_{E_4}
0.3	∞	∞	∞	0.462	0.462	∞	∞	∞	0.462	0.462	∞	∞	∞	0.462
0.4	∞	∞	∞	0.616	0.615	∞	2.98	4.14	0.615	0.615	∞	1.78	2.09	0.615
0.5	∞	2.95	4.14	0.827	0.826	∞	1.68	1.96	0.825	0.825	∞	1.34	1.53	0.825
0.6	∞	1.99	2.19	1.135	1.132	∞	1.46	1.59	1.132	1.131	∞	1.26	1.35	1.130
0.7	∞	2.06	2.02	1.638	1.628	∞	1.60	1.64	1.630	1.626	∞	1.42	1.47	1.626
0.8	∞	2.82	2.49	2.620	2.597	∞	2.17	2.14	2.601	2.592	∞	1.89	1.95	2.591
0.9	∞	5.82	4.36	5.506	5.441	∞	4.16	3.83	5.454	5.423	∞	3.84	3.83	5.423

Table 4: Optimal time limits as function of the load: Erlang E_2 service.

the behaviour of the optimal time limits as function of the load with that of the optimal Bernoulli schedules in [4]. In particular, the optimal set of time limits is, for each set of cost factors, such that at least one time-limit is infinite (i.e., the corresponding queue is served exhaustively), and the queues for which the time limits are infinite are the queues for which the ratio c_j/ρ_j is maximal over $j = 1, \ldots, s$ (in agreement with the '$c\mu$'-rule for priority systems). For the other queues it holds that the optimal time limit tends to infinity in light traffic ($\rho \downarrow 0$) as well as in heavy traffic ($\rho \uparrow 1$). In light traffic, finite time limits might force the server to make an often unnecessary tour along the queues to search for jobs which will only be present with small probability. In heavy traffic, the time limits have to be large in order to keep the system stable, i.e., to compensate for the loss of server availability due to the switching times.

In tabel 5 the optimal time limits are shown for various values of the equal upper bounds $U_h = U$, $h = 1, \ldots, P$, and of B, for a periodic four-queue system with visit order 1,2,3,4,3,2, and with the following parameters: arrival rates $\lambda_1 = \lambda_4 = 0.3$, $\lambda_2 = \lambda_3 = 0.1$, exponential service times with means $\beta_j = 1.0$, $j = 1, 2, 3, 4$, exponential switching times with means $\delta_h = 0.1$, $h = 1, \ldots, 6$, Erlang E_2 timers at queues 1 and 4, exponential timers at queues 2 and 3, and cost factors $c_1 = c_4 = 0.4$, $c_2 = c_3 = 0.1$. The stations are arranged in a row. The end stations have the same characteristics, and are more heavily loaded than the middle stations which also have the same characteristics. Due to this symmetry, $\tau_{h+3}^* = \tau_h^*$, $h = 1, 2, 3$, and $E\{W_{5-j}\} = E\{W_j\}$, $j = 1, 2$. Note that $c_1/\rho_1 > c_2/\rho_2$. If U is relatively small then the binding constraint is $\tau_1^* = U$, cf. (6.2), and τ_2^* and τ_3^* decrease with decreasing U, unless U becomes very small; then τ_2^* and τ_3^* increase until they reach their upper bound U. If B is relatively small then the binding constraint is $\sum_{h=1}^{6} \tau_h^* = B$, cf. (6.3).

Our final example concerns the cyclic polling system with five queues which has been considered in

U	B	τ_1^*, τ_4^*	τ_2^*, τ_5^*	τ_3^*, τ_6^*	C	$E\{W_1\}, E\{W_4\}$	$E\{W_2\}, E\{W_3\}$
∞	∞	∞	1.60	1.47	5.04	4.69	6.43
20.0	∞	20.00	0.97	0.89	5.20	4.91	6.36
20.0	40.0	18.93	0.60	0.47	5.22	4.76	7.03
20.0	30.0	14.39	0.37	0.24	5.31	4.83	7.22
20.0	20.0	10.00	0.00	0.00	5.49	4.96	7.59
10.0	∞	10.00	0.69	0.62	5.47	5.43	5.62
10.0	20.0	10.00	0.00	0.00	5.49	4.96	7.59
10.0	10.0	5.00	0.00	0.00	6.04	6.19	5.46
2.0	∞	2.00	2.00	2.00	7.69	9.03	2.32
2.0	10.0	2.00	1.57	1.43	7.69	9.01	2.41
2.0	5.0	2.00	0.25	0.25	7.69	8.88	2.93
2.0	2.0	1.00	0.00	0.00	10.12	12.04	2.45

Table 5: Optimal time limits for a four-queue model with constraints.

Borst et al. [6], tables III and IX. The arrival rates are $\lambda_1 = 0.35$, $\lambda_2 = \ldots = \lambda_5 = 0.10$, the service times are exponential with means $\beta_j = 1.0$, $j = 1, \ldots, 5$, and the switching times are exponential with means $\delta_2 = 0.10$, $\delta_j = 0.05$, $j = 1, 3, 4, 5$. Hence, $\rho = 0.75$ and $\Delta = 0.30$. The cost factor for station 1 is fixed, $c_1 = \lambda_1 = 0.35$. The cost factors for the other stations are equal, and are denoted by $c_{2-5} \doteq c_2 = \ldots = c_5$. Table 6 shows the optimal time limits and the minimal cost for the unconstrained optimization problem with exponentially distributed timers, for several values of c_{2-5}. For comparison, this table also contains the optimal job limits K_j^*, $j = 1, \ldots, 5$, and the corresponding minimal cost. A job limits places a maximum on the *number* of jobs which may be served during a visit of the server to a station. The optimal set of job limits can only be determined by enumeration of all, infinitely many, sets of job limits. In [6], table IX, the supposedly optimal sets of job limits have been found by a limited enumeration and on the basis of a conjecture which implies that at least one of the optimal job limits is infinite, which means that the corresponding queue is served exhaustively. This conjecture is supported by the numerical results on systems with Bernoulli schedules in [4]. In contrast with Bernoulli parameters (probabilities) with their finite range time limits have an infinite range. Therefore, numerical procedures for optimization of systems with unconstrained time limits will stop at some finite value for all time limits. In cases of very large values of a time limit we have compared the cost of the found set of finite time limits with the cost corresponding a similar set of time limits but where the stations with an originally large time limit are served exhaustively. The so obtained results also confirm the conjecture in [6]. In the example with $c_{2-5} = 0.20$ in table 6 the minimal cost of 3.94 attainable with exponentially distributed timers is larger than the minimal cost of 3.90 attainable with job limits. However, when we apply this set of mean time limits with Erlang distributed timers then the cost reduces to 3.89 with an Erlang E_4 distributed timer and to an estimated 3.86 with a constant time limit, cf. (4.7). The estimated minimal costs for the case of constant timers are indicated in table 6 in the column with the header $C^{(\mathrm{Det})}$. It seems that the larger flexibility in adjusting the time limits allows a lower minimal cost than that which is possible with job limits which are restricted to integer values. Table 7 shows the optimal sets of job limits and the corresponding minimal costs for the same system, but with the constraint $\sum_{j=1}^{s} K_j \leq 20$ on the total number of services that the server is allowed to perform during a cycle. These results are based on $M = 29$ terms of the power-series expansions; the estimated errors are in the order of 1%, much more than the estimated errors in the performance measures for the systems with time limits. Our results deviate in some cases from

c_{2-5}	τ_1^*	τ_2^*	τ_3^*	τ_4^*	τ_5^*	C	$C^{(\mathrm{Det})}$	K_1^*	K_2^*	K_3^*	K_4^*	K_5^*	C
0.01	∞	0.00	0.00	0.00	0.00	0.88	0.88	∞	1	1	1	1	0.88
0.05	∞	1.41	1.37	1.34	1.32	1.76	1.74	∞	2	2	2	2	1.76
0.10	∞	∞	∞	∞	∞	2.63	2.63	∞	∞	∞	∞	∞	2.63
0.20	2.83	∞	∞	∞	∞	3.94	3.86	3	∞	∞	∞	∞	3.90
1.00	0.34	∞	∞	∞	∞	10.47	10.40	1	∞	∞	∞	∞	10.74

Table 6: Optimal time limits and job limits for a five-queue model without constraint.

c_{2-5}	τ_1^*	τ_2^*	τ_3^*	τ_4^*	τ_5^*	C	$C^{(\mathrm{Det})}$	K_1^*	K_2^*	K_3^*	K_4^*	K_5^*	C
0.01	15.00	0.00	0.00	0.00	0.00	1.00	0.90	16	1	1	1	1	0.88
0.05	13.95	0.30	0.27	0.25	0.23	1.87	1.79	12	2	2	2	2	1.77
0.10	8.14	1.73	1.72	1.71	1.70	2.81	2.75	8	3	3	3	3	2.67
0.20	1.94	3.28	3.27	3.26	3.25	4.15	3.99	3	4	4	4	5	3.91
1.00	0.22	3.71	3.70	3.69	3.68	11.18	10.78	1	4	5	5	5	10.75

Table 7: Optimal time limits and job limits for a five-queue model with a constraint.

those reported in [6], table III. Appendix B contains more elaborate data on the costs as functions of the job limits. The case $c_{2-5} = 1.00$ which we have added reveals that it may be far from optimal in the constraint case to restrict the search for the set of optimal job limits to those sets for which $K_2 = \ldots = K_5$, although all parameters related to queues 2–5 are equal. In this particular case it would lead to a cost of 13.27, 23% more than the minimal cost of 10.74. Because the service times are exponentially distributed in this model, the mean of the maximal number of services per visit to queue j is $1 + \tau_j/\beta_j$, cf. (A.5), when a mean time limit τ_j is applied, $j = 1, \ldots, 5$. Therefore, we have determined optimal mean time limits for the case of exponential timers under the constraint $\sum_{j=1}^{s} \tau_j \leq 15$, for comparison with optimal sets of job limits. The results are displayed in table 7. We have also determined the optimal time limits for the case of Erlang E_2 distributed timers. The corresponding minimal costs are somewhat less than the minimal costs corresponding to the case of exponential timers. The estimated minimal costs for the case of constant timers, $C^{(\mathrm{Det})}$, are still somewhat higher than the corresponding minimal costs with job limits. This feature must be due to the randomness of the service times. Finally, note that the optimal service displines in the case $c_{2-5} = 0.01$ are extremal. In such a case it might be profitable to consider generalized service displines in which, e.g., some queues are only served every second cycle.

Numerical experiments indicate that the optimization problems considered in this section possess a unique solution. In some cases, the objective function is very flat near the optimum. Both properties have also been found in [4] for the unconstrained optimization problem with Bernoulli schedules as service disciplines.

When using the PSA for optimization purposes it is often a good strategy for reducing computation time to start the search with a moderate number of terms of the power-series expansions, and then to improve the approximated optimum by using more terms.

7 Conclusions

Numerical results with the PSA show that many phases of Erlang distributed timers may be needed for accurate evaluation of systems with constant timers. However, good approximations

for systems with constant timers can be obtained by extrapolation from results for systems with Erlang distributed timers with only a few phases.
When minimizing the cost of waiting by optimizing the values of the timers the use of a small number of phases often yields timer values with cost of waiting close to minimal when used for systems with constant timers. The latter is partly due to the flatness of the cost function near its minimum.
The computation scheme of the PSA for the model discussed in this paper is readily generalized to Markovian arrival processes at the stations, cf. Van den Hout & Blanc [10]. However, the required storage capacity for the coefficients of the power-series expansions strongly increases. In fact, (5.7) has to be multiplied by the product of the number of stages of the Markovian arrival processes at the various stations in this case.

References

[1] Baker, J.E., I. Rubin. Polling with a general-service order table, *IEEE Trans. Commun.* **COM-35** (1987), 283-288.

[2] Blanc, J.P.C. Performance evaluation of polling systems by means of the power-series algorithm, *Annals Oper. Res.* **35** (1992), 155-186.

[3] Blanc, J.P.C. Performance analysis and optimization with the power-series algorithm, in: *Performance Evaluation of Computer and Communication Systems*, eds. L. Donatiello, R. Nelson (Springer, Berlin, 1993), 53-80.

[4] Blanc, J.P.C., R.D. van der Mei. Optimization of polling systems with Bernoulli schedules, *Performance Evaluation* **22** (1995), 139-158.

[5] Blanc, J.P.C., R.D. van der Mei. Computation of derivatives by means of the power-series algorithm, *INFORMS J. on Comp.* **8** (1996), 45-54.

[6] Borst, S.C., O.J. Boxma, H. Levy. The use of service limits for efficient operation of multistation single-medium communication systems, *IEEE/ACM Trans. Networking* **3** (1995), 602-612.

[7] Coffman, E. Jr., I. Mitrani, G. Fayolle. Two queues with alternating service periods, in: *Performance '87,* eds. P.-J. Courtois, G. Latouche (North-Holland, Amsterdam, 1988), 227-239.

[8] Fricker, C., M.R. Jaïbi. Monotonicity and stability for periodic polling models, *Queueing Systems* **15** (1994), 211-238.

[9] Leung, K.K. Cyclic-service systems with nonpreemptive, time-limited service, *IEEE Trans. Commun.* **COM-42** (1994), 2521-2524.

[10] Van den Hout, W.B., J.P.C. Blanc. The power-series algorithm for Markovian queueing networks, in: W.J. Stewart (ed.) *Computations with Markov Chains,* (Kluwer, Boston, 1995), 321-338.

A Number of services during a visit

This appendix is concerned with a discussion of the distribution of the number of jobs that can be served during a visit to a station with a time limit, provided that sufficient jobs are present at that station. Let T denote the random variable which determines when the time limit expires, with mean τ. Further, let the random variables $S_1, S_2, \ldots$ be some generic service times during one visit, with distribution $B(t)$, mean β_1 and LST $\beta(\zeta)$. Finally, let M denote the maximal number of services during a visit. Assume that sufficient jobs are present. Then, m services can be performed during a visit if the mth service can start before the timer expires and if the timer has expired at the end of the mth service, i.e., for $m = 1, 2, \ldots$,

$$\Pr\{M = m\} = \Pr\{S_1 + \cdots + S_{m-1} < T,\ S_1 + \cdots + S_m \geq T\}. \tag{A.1}$$

By conditioning on the value of T this implies: for $m = 1, 2, \ldots$,

$$\Pr\{M = m\} = \int_0^\infty \Pr\{S_1 + \cdots + S_{m-1} < t,\ S_1 + \cdots + S_m \geq t\} d\Pr\{T \leq t\}. \tag{A.2}$$

By conditioning on the realisation of $S_1 + \cdots + S_{m-1}$ this leads to:

$$\Pr\{M = m\} = \int_0^\infty [B^{(m-1)*}(t) - B^{m*}(t)] d\Pr\{T \leq t\}, \quad m = 1, 2, \ldots. \tag{A.3}$$

The mean number of services during a visit can be determined as

$$E\{M\} = \int_0^\infty \sum_{m=1}^\infty m[B^{(m-1)*}(t) - B^{m*}(t)] d\Pr\{T \leq t\}. \tag{A.4}$$

In the special case of exponentially distributed service times the foregoing expression reduces to

$$E\{M\} = \int_0^\infty \sum_{m=1}^\infty m \frac{(t/\beta_1)^{m-1}}{(m-1)!} e^{-t/\beta_1} d\Pr\{T \leq t\} = 1 + \frac{\tau}{\beta_1}. \tag{A.5}$$

In the special case of exponentially distributed time limits (A.4) can be simplified to

$$E\{M\} = \sum_{m=1}^\infty \frac{m}{\tau} \int_0^\infty e^{-t/\tau} [B^{(m-1)*}(t) - B^{m*}(t)] dt = \sum_{m=1}^\infty m[\beta^{m-1}(1/\tau) - \beta^m(1/\tau)] = \frac{1}{1 - \beta(1/\tau)}. \tag{A.6}$$

Finally, in the special case of constant time limits (A.4) becomes

$$E\{M\} = \sum_{m=1}^\infty m[B^{(m-1)*}(\tau) - B^{m*}(\tau)]. \tag{A.7}$$

B Systems with job limits

This appendix contains some additional data for the five queue system with job limits discussed in section 6 and considered in [6]. Table 8 shows the mean waiting times at the various queues and the waiting cost for various values of the cost factors c_{2-5} for sets of job limits with either $K_1 = \infty$ or $K_2 = \ldots = K_5 = \infty$. Table 9 shows the same performance measures for sets of job limits satisfying $\sum_{j=1}^s K_j = 20$. With decreasing value of K_1 we have first increased K_5, then K_4, K_3,

etcetera. This seems to the best strategy when the factor c_{2-5} is relatively large. When this factor is relatively small then it is better to increase first K_2, then K_3, and so on (see the last two rows of table 9). Note that $E\{W_1\}$ increases with decreasing K_1, while the other mean waiting times show a zigzag behaviour for $K_1 \geq 8$, while they are decreasing with decreasing K_1 for $K_1 < 8$. We have also evaluated intermediate cases with $K_1 = \infty$ in the context of table 8 which revealed similar behaviour of the mean waiting times. In particular, it turned out that the cost function with $c_{2-5} = 0.05$ is very flat near its minimum, in the unconstrained case even more than in the constrained case.

job limits					mean waiting times					cost with $c_{2-5} =$				
K_1	K_2	K_3	K_4	K_5	W_1	W_2	W_3	W_4	W_5	0.01	0.05	0.10	0.20	1.00
∞	1	1	1	1	1.86	5.72	5.74	5.76	5.78	0.88	1.80	2.95	5.25	23.65
∞	2	2	2	2	2.29	4.72	4.76	4.80	4.84	0.99	1.76	2.71	4.63	19.92
∞	3	3	3	3	2.54	4.36	4.40	4.45	4.50	1.07	1.77	2.66	4.43	18.59
∞	4	4	4	4	2.68	4.18	4.22	4.27	4.34	1.11	1.79	2.64	4.34	17.95
∞	∞	∞	∞	∞	2.87	3.99	4.03	4.08	4.14	1.17	1.82	2.63	4.25	17.24
8	∞	∞	∞	∞	3.56	3.47	3.52	3.55	3.60	1.39	1.95	2.66	4.07	15.38
4	∞	∞	∞	∞	4.66	2.82	2.84	2.86	2.88	1.74	2.20	2.77	3.91	13.03
3	∞	∞	∞	∞	5.27	2.54	2.55	2.58	2.61	1.95	2.36	2.87	3.90	12.12
2	∞	∞	∞	∞	6.36	2.19	2.20	2.23	2.27	2.32	2.67	3.12	4.00	11.12
1	∞	∞	∞	∞	10.12	1.75	1.77	1.81	1.86	3.62	3.90	4.26	4.98	10.74

Table 8: Performance of a five-queue model with job limits without constraint.

job limits					mean waiting times					cost with $c_{2-5} =$				
K_1	K_2	K_3	K_4	K_5	W_1	W_2	W_3	W_4	W_5	0.01	0.05	0.10	0.20	1.00
16	1	1	1	1	1.87	5.72	5.73	5.75	5.78	0.88	1.80	2.95	5.25	23.62
15	1	1	1	2	1.96	6.20	6.22	6.24	3.52	0.91	1.80	2.90	5.12	22.86
14	1	1	2	2	2.07	6.81	6.82	3.80	3.84	0.94	1.79	2.85	4.98	21.99
13	1	2	2	2	2.20	7.62	4.17	4.20	4.24	0.97	1.78	2.79	4.82	21.00
12	2	2	2	2	2.37	4.66	4.70	4.73	4.77	1.02	1.77	2.71	4.60	19.69
11	2	2	2	3	2.49	4.81	4.84	4.87	3.80	1.05	1.79	2.70	4.54	19.20
10	2	2	3	3	2.61	4.95	4.98	3.85	3.88	1.09	1.80	2.68	4.45	18.58
9	2	3	3	3	2.85	5.09	3.89	3.91	3.96	1.17	1.84	2.68	4.37	17.85
8	3	3	3	3	3.10	3.91	3.94	3.98	4.00	1.24	1.88	2.67	4.25	16.92
7	3	3	3	4	3.38	3.80	3.83	3.86	3.47	1.33	1.93	2.68	4.17	16.14
6	3	3	4	4	3.73	3.62	3.65	3.34	3.37	1.45	2.00	2.70	4.10	15.28
5	3	4	4	4	4.15	3.39	3.13	3.16	3.18	1.58	2.10	2.74	4.03	14.32
4	4	4	4	4	4.58	2.89	2.90	2.92	2.95	1.72	2.19	2.77	3.94	13.27
3	4	4	4	5	5.23	2.59	2.60	2.63	2.60	1.93	2.35	2.87	3.91	12.25
2	4	4	5	5	6.35	2.22	2.24	2.23	2.27	2.31	2.67	3.12	4.01	11.17
1	4	5	5	5	10.12	1.77	1.77	1.81	1.86	3.61	3.90	4.26	4.98	10.75
15	2	1	1	1	1.95	3.51	6.23	6.24	6.26	0.90	1.79	2.90	5.13	22.91
1	5	5	5	4	10.12	1.75	1.77	1.81	1.88	3.61	3.90	4.26	4.98	10.75

Table 9: Performance of a five-queue model with job limits and a constraint.